Testing of detection tools for AI-generated text

Year: 2023 | *Citations: 317* | *Authors: Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba*

Abstract

Recent advances in generative pre-trained transformer large language models have emphasised the potential risks of unfair use of artificial intelligence (AI) generated content in an academic environment and intensified efforts in searching for solutions to detect such content. The paper examines the general functionality of detection tools for AI-generated text and evaluates them based on accuracy and error type analysis. Specifically, the study seeks to answer research questions about whether existing detection tools can reliably differentiate between human-written text and ChatGPT-generated text, and whether machine translation and content obfuscation techniques affect the detection of AI-generated text. The research covers 12 publicly available tools and two commercial systems (Turnitin and PlagiarismCheck) that are widely used in the academic setting. The researchers conclude that the available detection tools are neither accurate nor reliable and have a main bias towards classifying the output as human-written rather than detecting AI-generated text. Furthermore, content obfuscation techniques significantly worsen the performance of tools. The study makes several significant contributions. First, it summarises up-to-date similar scientific and non-scientific efforts in the field. Second, it presents the result of one of the most comprehensive tests conducted so far, based on a rigorous research methodology, an original document set, and a broad coverage of tools. Third, it discusses the implications and drawbacks of using detection tools for AI-generated text in academic settings.